Difficulty: Undergraduate Students

Multivariate calculus has always been a core part of any mathematics degree, but is often taught through computation or simplified assumptions instead of foundational abstractions. In this series of posts, I describe *a few* of the key concepts that were glanced over in my undergraduate studies, but carefully scrutinized during my graduate studies. I will assume that readers are familiar with the material in an introductory multivariate calculus class.

A good portion of the material below can be found in Chapter 3 of *Iterative solution of nonlinear equations in several variables* by J. M. Ortega and W. C. Rheinboldt. A more in-depth presentation can be found in Appendix A of *Lectures on Modern Convex Optimization* by A. Ben-Tal and A. Nemirovski

The Fréchet derivative: a robust standard

We start our series with the "canonical" definition of a derivative in higher dimensions. A function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ is said to have a **Fréchet** (or **F-**) **derivative at** $x \in \mathbb{R}^n$ for a given norm $\|\cdot\|$ if there exists a (unique) linear operator $\mathcal{A}_x^f : \mathbb{R}^n \mapsto \mathbb{R}^m$, called the F-derivative, that satisfies the relation

$$(\alpha) \qquad \lim_{\Delta \to 0} \frac{\left\| f(x+\Delta) - \left[f(x) + \mathcal{A}_x^f(\Delta) \right] \right\|}{\|\Delta\|} = 0$$

where the limit is taken over all subsequences $\{\Delta_n\} \subseteq \mathbb{R}^n$ tending to zero. Notice that this is a natural generalization of the one-dimensional case where we replace absolute value errors with norm errors.

Before giving more definitions, let us discuss some nuances and implications of the above definition.

- In general, the definition of $\mathcal{A}_x^f$ is non-constructive.
- $\mathcal{A}_x^f$ is linear in its f parameter, i.e., $\mathcal{A}_x^{\lambda(f_1+f_2)} = \lambda(\mathcal{A}_x^{f_1} + \mathcal{A}_x^{f_2})$.
- In the one-dimensional case ($n = m = 1$), the F-derivative and derivative coincide in the sense that $\mathcal{A}_x^f(\delta) = \delta f'(x)$ for $\delta \in \mathbb{R}$.
- The F-derivative is independent of different choices of the norm $\|\cdot\|$. This follows from the fact that for two norms $\|\cdot\|$ and $\|\cdot\|'$ in $\mathbb{R}^n$, there always exist constants $c_1 > 0$ and $c_2 \geq c_1$ such that
$$c_1\|x\| \leq \|x\|' \leq c_2\|x\| \quad \forall x \in \mathbb{R}^n.$$

In other works of literature, we might see the following variations and applications.

- The definition in (α) may be equivalently written as

$$\lim_{y \to x} \frac{\left\| f(y) - \left[f(x) + \mathcal{A}_x^f(y-x) \right] \right\|}{\|y - x\|} = 0$$

  where the limit is over all subsequences $\{y_n\}$ going to x. (*Prove this as a simple exercise*!)

- $\mathcal{A}_x^f(\Delta)$ may be written as $Df(x)[\Delta]$, $Df_x(\Delta)$, or $f'(x)\Delta$ to emphasize the dependence on f and x.

- In optimization theory, the term $f(x) + \mathcal{A}_x^f(\Delta)$ is often called the **first-order approximation of f at x.**

Once we have the above definition of a derivative, we can make the several follow-up definitions. The function f is **Fréchet** (or **F-**) **differentiable at** x if its F-derivative $\mathcal{A}_x^f$ exists. Consequently, the function f is **Fréchet** (or **F-**) **differentiable** or has **Fréchet** (or **F-**) **differentiability** if it is F-differentiable at all points in $\mathbb{R}^n$.

Some important properties that are unique to F-differentiability are as follows.

- If f is F-differentiable at x then f is **continuous** at x, i.e., $\lim_{\bar{x} \to x} f(\bar{x}) = f(x)$.

- The set $\{f(x) + \mathcal{A}_x^f(\Delta) : \Delta \in \mathbb{R}^n\}$ is a tangent plane of f at x.

Finally, some important anti-properties of F-derivatives and F-differentiability are as follows.

- The subsequences $\{\Delta_n\}$ **need not** lie on a line, e.g., like in the definition of a **partial derivative** $\partial f_i / \partial x_j$.

- In fact, the existence of the partial derivatives at x **is generally not** sufficient to conclude F-differentiability at x.

  - *Exercise.* Consider the function

$$(\gamma_1) \qquad f(x_1, x_2) = \begin{cases} x_1, & \text{if } x_2 = 0, \\ x_2, & \text{if } x_1 = 0, \\ 1, & \text{otherwise}, \end{cases}$$

    at zero. Show that the partial derivatives of f exist at zero, but the function itself is not F-differentiable.

Difficulty: Undergraduate Students

Often, we do not need full F-differentiability of a multivariate function $f : \mathbb{R}^n \mapsto \mathbb{R}$ to derive interesting results about f. Below, we describe a weaker notion of differentiability and its intriguing properties.

The Gateaux derivative: provably weaker, but more flexible

The next stop on our journey is a weaker, but related notion of a derivative. A function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ is said to have a **Gateaux** (or **G-**) **derivative at** $x \in \mathbb{R}^n$ for a given norm $\|\cdot\|$ if there exists a (unique) linear operator $\mathcal{B}_x^f : \mathbb{R}_+ \mapsto \mathbb{R}^m$, called the G-derivative, that satisfies the relation

$$(\beta) \qquad \lim_{t \downarrow 0} \frac{\left\| f(x + t\Delta) - \left[f(x) + \mathcal{B}_x^f(\Delta) \right] \right\|}{t} = 0 \quad \forall \Delta \in \mathbb{R}^n$$

where the limit is taken over all positive subsequences $\{t_n\} \subseteq \mathbb{R}_+$ tending to zero.

Like in the previous post, let us discuss a few nuances and implications of the above definition.

- In general, the definition of $\mathcal{B}_x^f$ is non-constructive.
- $\mathcal{B}_x^f$ is linear in its f parameter, i.e., $\mathcal{B}_x^{\lambda(f_1+f_2)} = \lambda(\mathcal{B}_x^{f_1} + \mathcal{B}_x^{f_2})$.
- Compared to the F-derivative, the subsequences in the G-derivative are restricted to subsequences $\{t_n d\}$ which lie on a line emanating from the origin.
- In the one-dimensional case ($n = m = 1$), the F-derivative and G-derivative coincide.
- Similar to the F-derivative, the G-derivative is independent of different choices of the norm $\|\cdot\|$ (for the same reasons).

Some works may have the same notation for the Fréchet and Gateaux derivatives, while others may prefer $D_x f(x)[\Delta]$ for Fréchet and $Df(x)[\Delta]$ for Gateaux. The two notions may be related as follows.

- If f has an F-derivative $\mathcal{A}_x^f$ at x then (i) it also has a G-derivative $\mathcal{B}_x^f$ at x, and (ii) $\mathcal{A}_x^f = \mathcal{B}_x^f$.
- If f is convex, then the F-derivative $\mathcal{A}_x^f$ exists at x if and only if the G-derivative $\mathcal{B}_x^f$ exists at x.

We now make the corresponding follow-up definitions. The function f is **Gateaux** (or **G-**) **differentiable at** x if its G-derivative $\mathcal{B}_x^f$ exists. Consequently, the function f is **Gateaux** (or **G-**) **differentiable** or has **Gateaux** (or **F-**) **differentiability** if it is G-differentiable at all points in $\mathbb{R}^n$.

With the above definitions in mind, we give a few properties that merely require G-differentiability (instead of F-differentiability).

- If f is G-differentiable at x, then f is **hemicontinuous** at x, i.e., for any $\varepsilon > 0$ and $\Delta \in \mathbb{R}^n$ there exists $\delta = \delta(\varepsilon, \Delta)$ such that whenever $|t| < \delta$ then $\|f(x+t\Delta) - f(x)\| < \varepsilon$.
  - *Exercise.* Consider the function

    $$(\gamma_2) \qquad f(x_1, x_2) = \begin{cases} 0, & \text{if } x = 0, \\ \dfrac{x_2(x_1^2 + x_2^2)^{3/2}}{(x_1^2 + x_2^2)^2 + x_2^2}, & \text{otherwise.} \end{cases}$$

    Show that f has a G-derivative at zero, but not an F-derivative at zero. Show, moreover, that the G-derivative is hemicontinuous at zero.

- If f is G-differentiable at x and $\|x\| = x^T x$ for every $x \in \mathbb{R}^n$, then the partial derivatives of f exist at x. Furthermore, the matrix representation of $\mathcal{B}_x^f$ is given by the **Jacobian**

  $$\mathcal{B}_x^f \equiv \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x) & \cdots & \frac{\partial f_1}{\partial x_m}(x) \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1}(x) & \cdots & \frac{\partial f_n}{\partial x_m}(x) \end{pmatrix} \in \mathbb{R}^{n \times m};$$

  its transpose in the case of $m = 1$ is called the **gradient of** f **at** x, and is denoted by $\nabla f(x) \in \mathbb{R}^n$.
  - More general versions of the Jacobian and gradient exist for other inner product spaces through the use the **Riesz-Fréchet representation** theorem. The (generalized) gradient is specifically defined as the unique element $\nabla f(x) \in \mathbb{R}^{n\times m}$ satisfying

    $$\left\langle (\mathcal{B}_x^f)^* \tau, \Delta \right\rangle = \langle \nabla f(x)\tau, \Delta \rangle$$

    for every $\tau \in \mathbb{R}^m$ and $\Delta \in \mathbb{R}^n$, where $(\mathcal{B}_x^f)^*$ is the adjoint operator of $\mathcal{B}_x^f$.

- [**Chain Rule**] If $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ has a G-derivative at x and $g : \mathbb{R}^m \mapsto \mathbb{R}^p$ has an F-derivative at $f(x)$, then the composite function $h := g \circ f$ has a G-derivative at x where

  $$\mathcal{B}_x^h = \mathcal{A}_{f(x)}^g \mathcal{B}_x^f.$$

  If, in addition, $\mathcal{B}_x^f$ is an F-derivative then $\mathcal{B}_x^h$ is an F-derivative as well.
  - *Exercise.* Let f be as in (γ_2) and let $g : \mathbb{R}^2 \mapsto \mathbb{R}$ be given by $g(x) = (x_1, x_2^2)^T$. Show that the composite function $h = f \circ g$ does not have a G-derivative at zero.

Finally, some important anti-properties of G-derivatives and G-differentiability are as follows.

- Surprisingly, just like the F-derivative, the existence of the partial derivatives at x **do not** imply that the G-derivative exists at x.
  - *Exercise.* Consider the same function in equation (γ_1) of the previous post. Show that f is not G-differentiable at zero.
- The G-differentiability of f **does not** imply that f is continuous (unlike for F-differentiability).
  - *Exercise.* Consider the function

$$f(x_1, x_2) = \begin{cases} 0, & \text{if } x_1 = 0, \\ \dfrac{2x_2 \exp(-x_1^{-2})}{x_2^2 + \exp(-2x_1^{-2})}, & \text{otherwise,} \end{cases}$$

    at zero. Show that the G-derivative of f exists at zero, but f is not continuous.

Difficulty: Undergraduate Students

Sometimes notions in one-dimensional spaces are not easy to generalize to multi-dimensional spaces. I believe the notion of a differential (or the derivative in one-dimensional space) is one of them.

Differentials: a useful, but naive surrogate

In one-dimensional calculus, students are typically introduced to the definition of the derivative of a function $f : \mathbb{R} \mapsto \mathbb{R}$ by constructive means. Specifically, if

$$(\theta) \qquad \lim_{t \downarrow 0} \frac{f(x+t)-f(x)}{t} = \lim_{t \uparrow 0} \frac{f(x+t)-f(x)}{t} = \lim_{t \to 0} \frac{f(x+t)-f(x)}{t} = f'(x)$$

then $f'(x)$ is called the (univariate) **derivative of f at x**. In this special setting, the existence of $f'(x)$ enjoys all the nice properties of the F-derivative (e.g., continuity and tangency) and G-derivative (e.g., chain rule) *without the issues of construction*.

It is then natural to ask whether (θ) could be extended to the multivariate setting while simultaneously keeping (i) its constructive nature *and* (ii) the nice properties of the F-derivative (and G-derivative). Below, we show an approach of obtaining (i) which partially obtains (ii).

A function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ is said to have a **Gateaux** (or **G-**) **differential at** $x \in \mathbb{R}^n$ in the direction $\Delta \in \mathbb{R}^n$ if the function

$$(\pi_1) \qquad V^f(x, \Delta) = \lim_{t \to 0} \frac{f(x+t\Delta)-f(x)}{t},$$

called the G-differential, is well-defined. Here, the limit is taken over all subsequences $\{t_n\} \subseteq \mathbb{R}$. If, in addition,

$$(\pi_2) \qquad \lim_{\Delta \to 0} \frac{\|f(x+t\Delta)-f(x)-V^f(x,\Delta)\|}{\|\Delta\|} = 0$$

where the limit is taken over all subsequences $\{\Delta_n\} \subseteq \mathbb{R}^n$, then f is also said to have a **Fréchet** (or **F-**) **differential at** $x \in \mathbb{R}^n$ (also denoted by $V^f(x, \Delta)$).

Let us now make a few glancing remarks.

- Unlike the definitions of an F-derivative or G-derivative, the definition of V^f in (π_1) is constructive.
- Similar to the one-dimensional case, the well-definedness of $V^f(x, \Delta)$ equivalent to the left and right limits (in terms of t) being equal.

- In the one-dimensional case of $n = m = 1$, we have

$$V^f(x, 1) = f'(x) = -V^f(x, -1).$$

- It is straightforward to see that if $V^f(x, \Delta)$ exists for $\Delta \in \mathbb{R}^n$ and is linear in Δ then $V^f(x, \Delta) = \mathcal{B}_x^f(\Delta)$. Furthermore, if (π_2) holds then $V^f(x, \Delta) = \mathcal{A}_x^f(\Delta)$.

While we have fulfilled property (i), the following anti-properties (given as exercises) show that property (ii) cannot be fully realized.

- *Exercise.* Consider the function

$$f(x_1, x_2) = \begin{cases} 0, & \text{if } x = 0, \\ \dfrac{x_1 x_2^2}{x_1^2 + x_2^4}, & \text{otherwise.} \end{cases}$$

  Show that $V^f(0, \Delta)$ exists for every $\Delta \in \mathbb{R}^2$, but f does not have a G-derivative at zero. As a bonus, show that f is not continuous at zero.

- Consider the function

$$f(x_1, x_2) = \operatorname{sgn}(x_2) \min(|x_1|, |x_2|)$$

  which is clearly continuous at zero. Show that $V^f(0, \Delta)$ exists for every $\Delta \in \mathbb{R}^2$, but f does not have a G-derivative at zero.

On the other hand, the nice property about the G-differential (resp. F-differential) is that it gives a good initial estimate for the G-derivative (or F-derivative) if one can extract the appropriate linear form from it (and verify property (π_2) in the case of the F-derivative) or we know that f is G-differentiable (resp. F-differentiable). A classic example of the utility of the differential is in deriving the derivative of the log determinant of a matrix, which we show below.

Let $f : \mathcal{S}_{++}^n \mapsto \mathbb{R}$ be given by $f(M) = \log \det M$, where $\mathcal{S}_{++}^n$ denotes the space of positive definite matrices. Since $\log(\cdot)$ is differentiable on $\mathbb{R}_{++}$ and $h(M) = \det(M)$ is a polynomial function of the components of M (and, hence differentiable) we can obtain the derivative of f by differentials and the chain rule. Using some standard linear algebra techniques, the F-differential of h (and, hence, the F-derivative of h) is then given by

$$\begin{aligned} V^h(M, \Delta) &= \lim_{t \to 0} \frac{\det(M + t\Delta) - \det(M)}{t} \\ &= \lim_{t \to 0} \frac{\det(M[I + tM^{-1}\Delta]) - \det(M)}{t} \\ &= \det(M) \lim_{t \to 0} \frac{\det(I + tM^{-1}\Delta) - 1}{t} \\ &= \det(M)\operatorname{tr}(M^{-1}\Delta) = \mathcal{A}_M^h(\Delta), \end{aligned}$$

Hence, by the chain rule, we have

$$\mathcal{A}_M^f(\Delta) = \mathcal{A}_{\det M}^{\log(\cdot)} \mathcal{A}_M^h(\Delta) = \frac{\det(M)\mathrm{tr}(M^{-1}\Delta)}{\det(M)} = \mathrm{tr}(M^{-1}\Delta).$$

One can even obtain the gradient $\nabla f(x) = M^{-T}$ by using the fact that

$$\mathcal{A}_M^f(\Delta) = \mathrm{tr}(M^{-1}\Delta) = \left\langle M^{-T}, \Delta \right\rangle.$$

To close, let us present this nice schematic of the various relations between G/F-derivatives and G/F-differentials in terms of the displacement variable Δ that shows up in $\mathcal{A}_x^f(\Delta)$, $\mathcal{B}_x^f(\Delta)$, and $V^f(x, \Delta)$.